CIND820-Big Data Analytics Final Report

A Study on Predicting Strokes

Name: Ibrahim Sayed Ahmed

Student Number: 501120722

Ryerson University

Name: Ibrahim Sayed Ahmed
Student Number: 501120722

Table of Contents

Name: Ibrahim Sayed Ahmed
Student Number: 501120722

1. Abstract

Introduction:

Stroke, also known as brain attack, happens when blood flow to the brain is blocked, preventing it from getting oxygen and nutrients from it and causing the death of brain cells within minutes. According to the World Health Organization (WHO), it is the second cause of death worldwide after ischemic heart disease. Stroke victims can experience paralysis, impaired speed, or loss of vision. While some of the Stroke risk factors cannot be modified, such as family history of cerebrovascular diseases, age, gender and race, others can and are estimated to account for 60% -80% of stroke risk in the general population. Therefore, predicting stroke outcome for new cases can be determining for them to be treated early enough and avoid disabling and mortal consequences. The purpose of this research is to create an accurate model for predicting Stroke outcome with Data Science and Machine Learning (ML) based on previous data and individual characteristics. Providing useful information for the medical staff to deploy the needed treatment and decrease risks and consequences.

Objective:

The purpose of this project is to apply Classification and Predictive Analysis on the data:

- Identify the most important factors for stroke prediction.
- Visualize the results
- Propose a predictive analytics approach for stroke prediction.

Research Question:

I. What are the attributes related to stroke?
II. What is the correlation between pre-existing health conditions and stroke risks?
III. How to predict stroke risk?

The plan is to use data collected to identify the most important factors for stroke prediction.

The machine learning methods, techniques to be applied to address these questions are as follows:

- ❖ What are the attributes related to stroke?

Using Exploratory Data Analysis

- ❖ Implementation of supervised ML classification Algorithms

Building several models and compare which will be the best, Like Decision Tree Classifier (DT), Random Forest Classifier (RF), K-Nearest Neighbour (KNN)

2. Literature Review

This study looked at how machine learning techniques were used on stroke patients to predict stroke, which will also help to understand and resolve the problem in more effective ways.

This section analyses and reviews the previously published papers that deal with the work on stroke disease prediction using different machine learning approaches and algorithms.

José Alberto [1] proposed Stroke prediction through Data Science and Machine Learning Algorithms. He used classification algorithms-decision Tree, Naive Bayes, Logistic regression, random forest, SVM, Neutral Network and others for predicting the stroke with related attributes. He compared these techniques and chose Random Forest to be the best performing ML algorithm for this specific problem. As for many models for disease prediction, the cost of having false negatives is much higher than the one if predicting false positives. This is because, even though the person is not prone to suffering from a stroke, taking precaution and the recommended treatment has low negative impact in person's health. On the other hand, if we suppose that we say that the person has not risk of having such disease, we can even put its life in danger. Therefore, it is very important for the model to be evaluated depending on the Business problem to be solved. The model resulting from this research shown a very low rate of false negative predictions, so it seems that this model works with good performance.

Jenna et al. [2] provides a study of various risk factors to understand the probability of stroke. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. She used Support Vector Machine (SVM). In this work, we have implemented SVM with different kernel functions and found that linear kernel gave an accuracy of 90 %.

A predictive analytics approach for stroke prediction using machine learning and neural networks research [3] show using machine learning and neural networks in the proposed approach. They performed feature correlation analysis and a step wise analysis for choosing an optimum set of features. Three machine learning algorithms were implemented on a set of different features and principal components configurations. They found that neural network works the best. The accuracy and miss rate for this combination are 78% and 19% respectively.

Sudha. A [4] under the guidance of her professors N. Jaisankar & P Gayathra, proposed a stroke predictive Model using classification techniques. They used classification algorithms-decision Tree, Naive Bayes & Neural Networks for predicting the stroke with related attributes. They utilized principal component analysis algorithm for dimension reduction. They studied & used sensitivity 7 accuracy indicators for evaluation. Decision tree achieved 95.29% of sensitivity & 98.01% of accuracy. Bayesian classifier achieved 8710% & 91.30% respectively. They compared these techniques and chose the decision tree as the best classification Method.

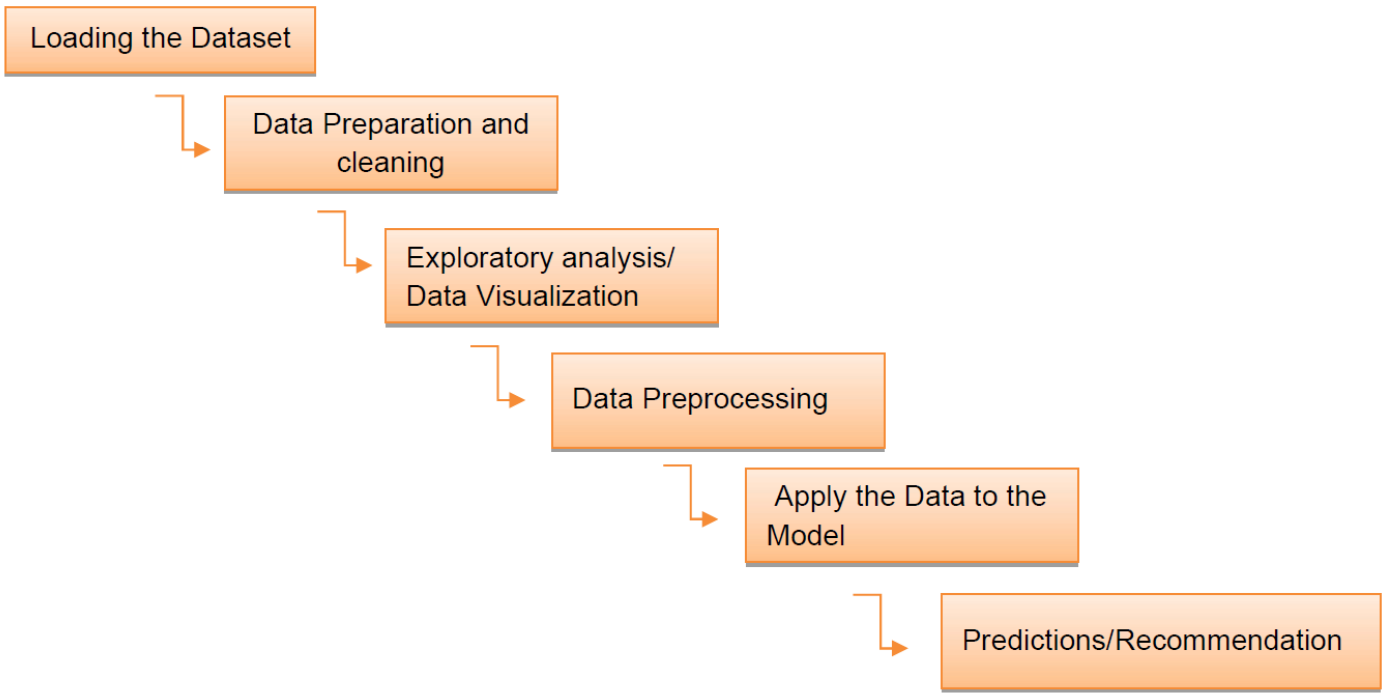
Department of Computer Architecture [5] and Automation team of Universidad Completeness de Madrid, Spain along with hospital Universitario de La Princesa, Madrid, Spain researched about the testing the hypothesis that state of art machine learning based modeling methods. They investigated if non-invasive monitoring technologies could aid in stroke type diagnosis. These methods can even be used to predict risks in the future, such as the patient's ultimate mortality. With 7 predictors and 2 goal variables—prediction of stroke type and prediction of death—they gathered a dataset made up of the medical records of 119 patients. They assessed over 6 distinct metrics using 7 different machine learning algorithms, including Decision Tree, KNN, Logistic Regression, Naive Bayes, Neural Network, Random Forest, and Support Vector Machines. Additionally, they used the 10-fold cross validation re-sampling technique for the trained classifier's validation against any unknown sample and a guaranteed validation set from training one. Sensitivity, specificity, accuracy, the F measure, and areas under ROC as well as PRC were the model performance evaluation metrics utilized to compare the algorithms. With values of 0.93+0.03 and 0.970.01, respectively, Random Forest models produced the greatest results overall in the diagnosis of stroke prediction and death prediction when compared to other algorithms.

A team of National Institute of Engineering, Karnataka, conducted a survey about AI applications in stroke and aimed to predict the accurate results of occurrence of stroke. They used predictive algorithms and parameters that include patients’ characteristics like gender, age, height, BMW, etc., they built a data model using decision tree algorithm to analyze these parameters. The result was analyzed using confusion matrix and the accuracy was 95%. To achieve this, they built the training model that helped to compare the newly fed data with the survey data. And the report was generated based on this comparison.

The outcomes of the various techniques show that a variety of factors can have an impact on the results of any study that has been performed. Any study's conclusion will be influenced by these numerous aspects, which include the method of data collection, the features used, the method used to clean the data, the method used to impute missing values, randomization, and standardization of the data. It is crucial that the researchers understand the relationships between the different input factors in an

electronic health record and how they affect the final stroke prediction accuracy.

3. Approach Overview



4. Overview of the Data

The dataset to be used for this project are Stroke Prediction Dataset.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

The dataset has about 5k records. There are 12 columns: -id - gender - age - hypertension -heart_disease - ever married - work_type - Residence_type - avg_glucose_level - bmi -smoking_status – stroke.

Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if not

Loading Packages and Data Preparation & Cleaning

Python was the best option when deciding which programming language to use for this project since it was the most adaptable and offered a wide range of packages for both basic statistical analysis and the building of machine language models.

In order to prepare, build, and plot the dataset, I first loaded the libraries Numpy, Pandas, Sklearn, and Mathplotlib.

Displaying the columns, rows, and overall structure of the data frame is the best approach to make sense of the dataset. Using the Pandas package, the initial loading and cleaning were carried out.

I did the following after importing the dataset into the variable "data":

- **data.head():** resulted in the first few columns and rows

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

In [4]:
```
df.describe()
```

Out[4]:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

In [6]:
```
df['stroke'].value_counts()
```

Out[6]:
```
0    4861
1     249
Name: stroke, dtype: int64
```

this dataset was highly unbalanced. Only 249 patients suffered a stroke while the remaining 4861 patients did not have the experience.

df.isnull().sum(): show missing values in data, bmi column had 201 missing values, so records with empty value in BMI was replaced with mean of BMI.

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

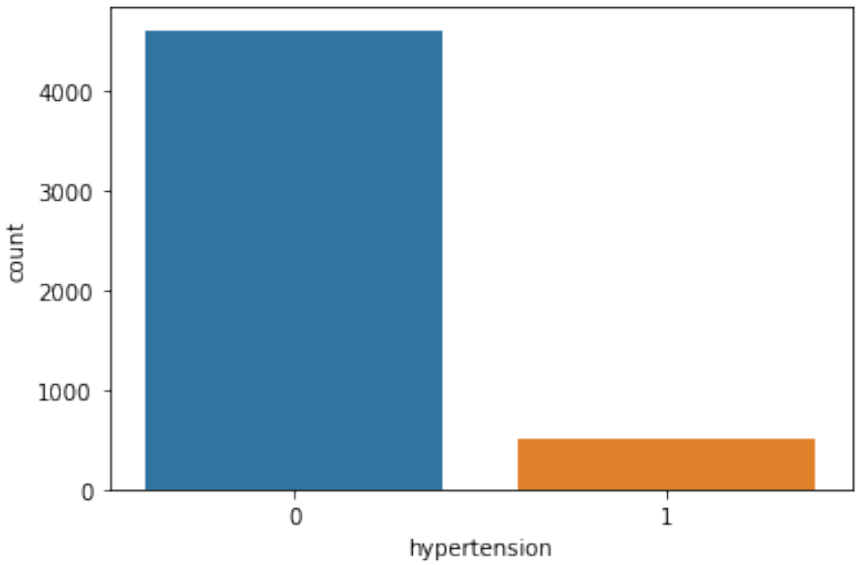
5. **EDA (Exploratory Data Analysis)**

Is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Visualization

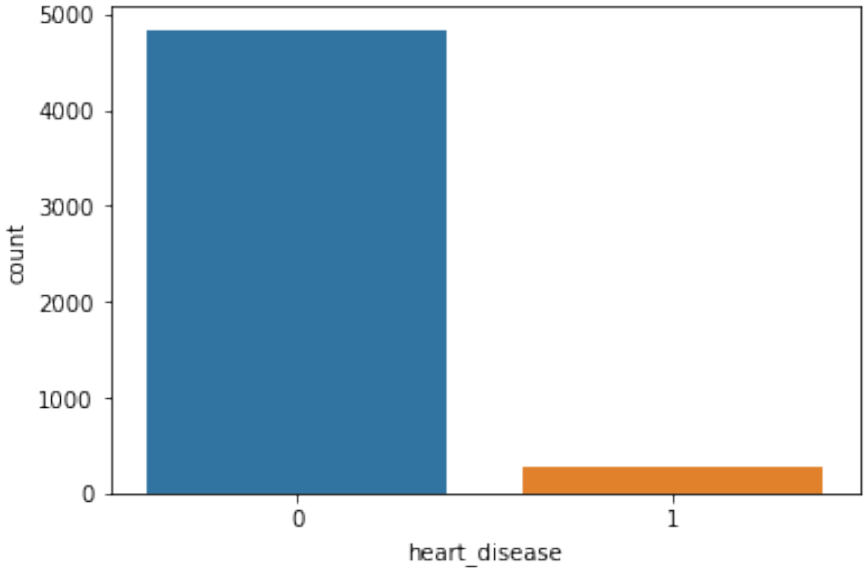
Hypertension attribute

498 patients with hypertension which represents at round 10 % of the sample.



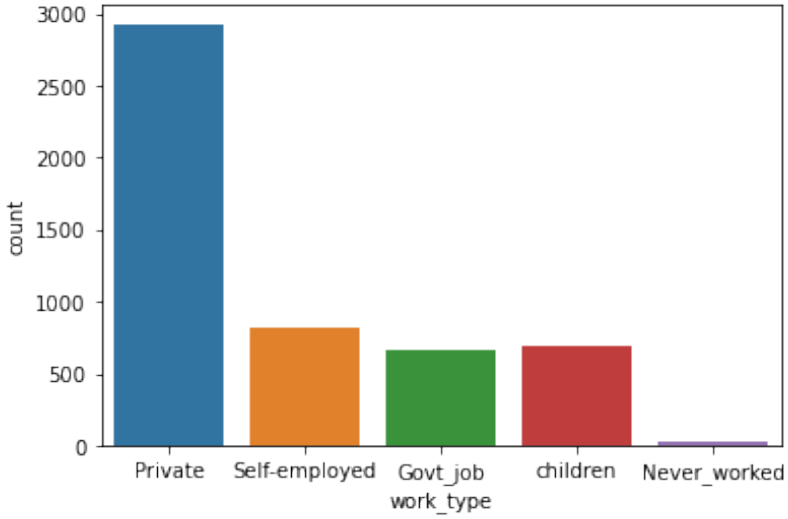
Heart disease attribute

276 patients with heart disease which is 5.4 % of the sample.



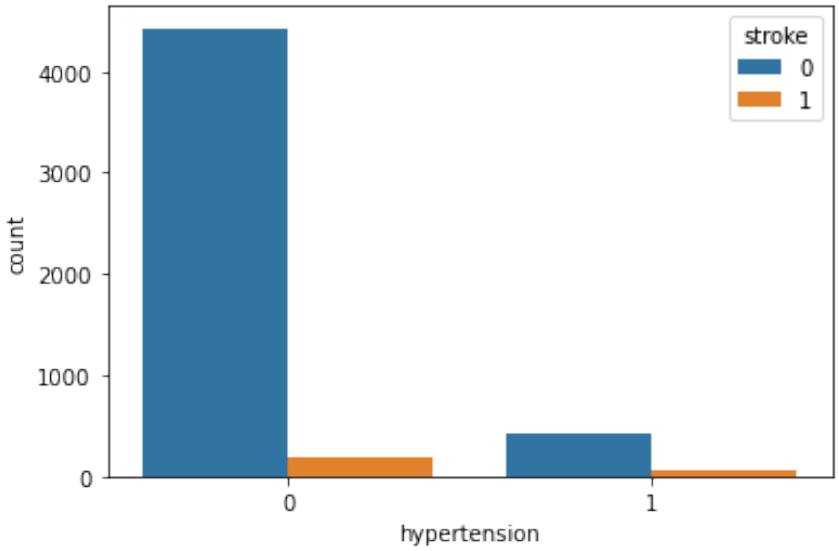
Work type attribute

- 2925 people work in the private sector.
- 819 people are self-employed
- 657 people work at the government job.



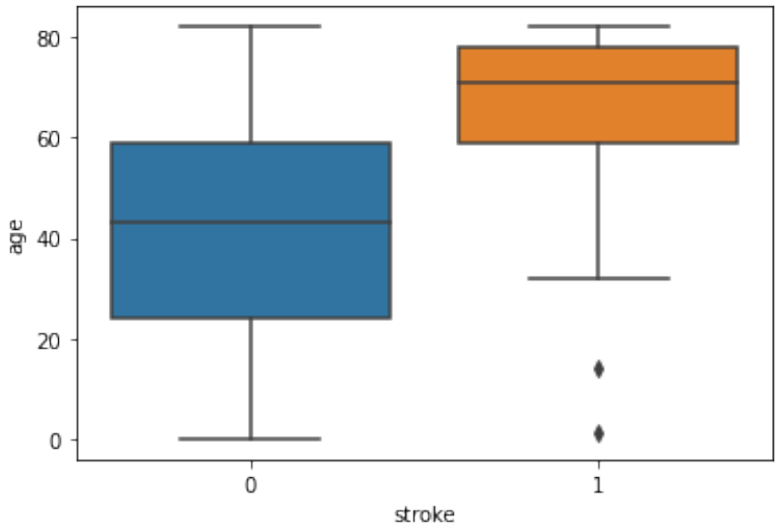
Hypertension affects on stroke

Young people rarely have hypertension, while the elderly frequently does. A stroke can be brought on by hypertension. Our data do not paint a very clear picture of hypertension. On patients with hypertension, there is not a lot of information.



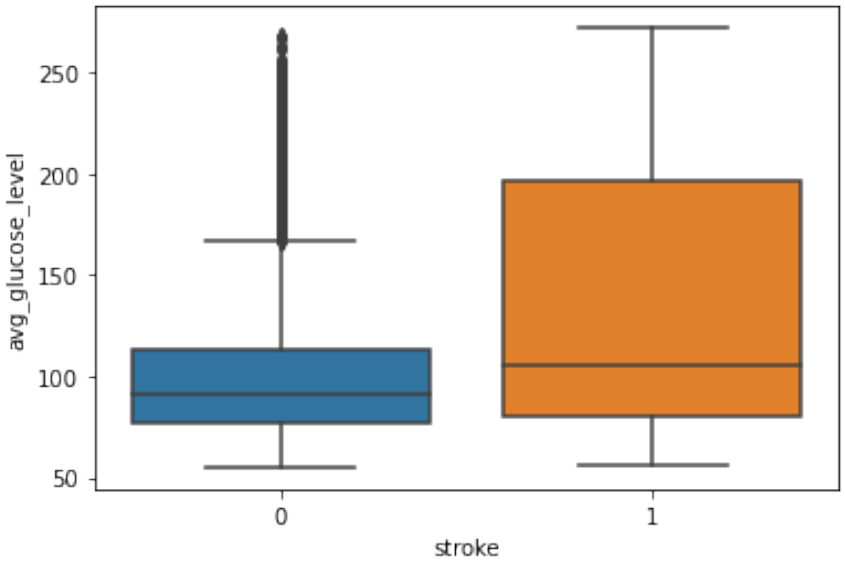
Age effects on stroke

People over 60 have a higher risk of having a stroke. Some irregularities can be attributed to strokes in people under the age of twenty. Given that our lifestyle choices and eating habits influence stroke risk, the information is probably accurate. Another conclusion is that adults above the age of 60 comprise the group of people who do not have strokes.



Avg glucose level

People having stroke have an average glucose level of more than 100.



Correlation matrix

- We can confirm that some of the variables are multicollinear based on the correlation matrix mentioned above. An example of this is the 0.68 correlation between the ever married and age columns.

6. Data Processing & Modeling

Classification Algorithm

Supervised learning algorithms like decision trees, random forests, Support Vector Machine Classification, Logistic Regression, K Nearest Neighbour and naive Bayes algorithms are evaluated to check for the features that contribute to predicting stroke after going through the available techniques used for classification for early stroke detection.

Decision Tree

A model is developed using a supervised learning algorithm to predict the class of a target variable based on the target set's features. The algorithm assumes that each feature's presence or absence in the dataset depends on the other features and helps to assign the target to a particular class.

A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically. [9]

Random Forest

Classification and regression problems can be resolved using ensemble learning techniques like Random Forests (sometimes referred to as random choice forests). They work through distributed training of many decision trees. A random forest's output is the class chosen by most trees while dealing with classification issues. When it comes to the specifics of ensemble learning, which is the practice of using multiple classifiers to solve complicated problems and increase the performance of a model, it is built. According to the Random Forest classifier's name, "combining a large number of decision trees on different subsets of a given dataset and taking an average to enhance the projected accuracy" The random forest considers the forecasts from all the trees as opposed to just one. The most popular forecasts are used to determine the outcome.

Support Vector Machine Classification

Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

A simple linear SVM classifier works by making a straight line between two classes. That means all the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Naive Bayes Classification

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, some fruit may be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

K-Nearest Neighbors Algorithm

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing K, the user can select the number of nearby observations to use in the algorithm.

K is the number of nearest neighbors to use. For classification, a majority vote is used to determine which class a new observation should fall into. Larger values of K are often more robust to outliers and produce more stable decision boundaries than very small values (K=3 would be better than K=1, which might produce undesirable results.

Logistic Regression

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

Results

In this study, different categorization algorithms were used to make early stroke predictions.

Logistic Regression, Random Forest Classification and Support Vector Machine models achieved the highest accuracy rate among all the classification algorithms.

Conclusion

According to this study's findings, it is possible to forecast stroke prediction using machine learning algorithms. Well-known classification algorithms, such as Decision Tree, Support Vector Machine, K Nearest Neighbour, Logistic Regression, Random Forest Classification and Support Vector Machine were tested for their accuracy rates.

Not all strokes can be prevented by changing one's lifestyle. Even while some of these adjustments can greatly reduce your risk of stroke, they can all be highly advantageous. If the person immediately stops smoking, the risk of stroke will be lowered. There should be restrictions on alcohol. Alcohol can raise blood pressure, which ups the chance of suffering a stroke. Maintain a healthy weight. Being overweight or obese puts a higher risk of having a stroke. Eating a balanced diet and exercising frequently will help maintain a healthy weight. It needs to take all these precautions to avoid a stroke.

A link to repository on GitHub

https://github.com/Ibrahim-S-Ahmed/Stroke_Prediction_Capston_Project

References

1. https://www.researchgate.net/publication/352261064_Stroke_prediction_through_Data_Science_and_Machine_Learning_Algorithms
2. Stroke prediction using SVM | IEEE Conference Publication | IEEE Xplore
3. https://www.sciencedirect.com/science/article/pii/S2772442522000090
4. https://www.semanticscholar.org/paper/Effective-Analysis-and-Predictive-Model-of-Stroke-Sudha-Gayathri/d46f2f63cc3ca9714c767b805cd033776f0bd3ee
5. https://www.researchgate.net/publication/335515560_Comparison_of_Different_Machine_Learning_Approaches_to_Model_Stroke_Subtype_Classification_and_Risk_Prediction
6. https://www.hindawi.com/journals/bn/2022/7725597/
7. https://www.ijraset.com/research-paper/automated-prediction-of-brain-stroke-disease-classification
8. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234722
9. https://www.lucidchart.com/pages/decision-tree